Published in final edited form as:
Statistics in Medicine (June 2017), **36**(13): 2016-2031
DOI: 10.1002/sim.7269

A two-phase model for chronic disease processes under intermittent inspection

Ying Wu
Institute of Statistics,
Nankai University

Richard J. Cook
Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: rjcook@uwaterloo.ca

Summary

A model is developed for chronic diseases with an indolent phase that is followed by a phase with more active disease resulting in progression and damage. The time scales for the intensity functions for the active phase are more naturally based on the time since the start of the active phase, corresponding to a semi-Markov formulation. This two-phase model enables one to fit a separate regression model for the duration of the indolent phase and intensity-based models for the more active second phase. In cohort studies for which the disease status is only known at a series of clinical assessment times, transition times are interval-censored, which means the time origin for phase II is interval-censored. Weakly parametric models with piecewise constant baseline hazard and rate functions are specified, and an expectation-maximization algorithm is described for model fitting. Simulation studies examining the performance of the proposed model show good performance under maximum likelihood and two-stage estimation. An application to data from the motivating study of disease progression in psoriatic arthritis illustrates the procedure and identifies new human leukocyte antigens associated with the duration of the indolent phase.

Keywords: expectation-maximization algorithm; interval-censoring; recurrent events; two-phase process; two-stage estimation

1 Introduction

1.1 Disease Processes with Delayed Activity

Many chronic disease processes feature considerable variability in their course which must be dealt with in statistical analysis for valid inference. Regression modeling and regression diagnostics play a

central role in explaining this variation in such a way that scientific understanding can be advanced. Another avenue is to generalize the family of stochastic models considered as the basis for analysis. Finite mixture models, for example, offer an appealing generalization as they involve conceptualization of two or more subpopulations of individuals, each with different stochastic models generating the response process. When models are directed at dynamic aspects of disease processes the simplest and perhaps most studied mixture model accommodates a non-susceptible sub-population of individuals whose status will never change, while individuals in the complementary sub-population experience the disease process. Such models are often called cure-rate models when modeling the time to an event (Farewell, 1986) or mover-stayer models when considering multistate disease processes (Goodman, 1961, Frydman, 1984).

In many contexts it is unnatural to envision diseased individuals as being indefinitely at zero risk of disease activity or damage. An alternative, and less extreme assumption is to consider two phases of the disease course: an indolent phase (phase I) during which affected individuals do not experience clinically meaningful disease activity or damage and an active phase II of disease progression. Chronic diseases whose course can be represented in this way include HIV/AIDS where phase I represents the period of HIV infection prior to the manifestation of AIDS defining events, and phase II represents the onset of opportunistic infections or death. In diabetes there may be a long phase I period during which no symptoms are evident, followed by a second phase during which there is evidence of retinopathy, nephropathy or other types of vascular impairment. Individuals with hepatitis C infection may go a long time without experiencing any liver cirrhosis but will ultimately experience progressive liver damage. Finally arthritis patients may simply have elevated markers of inflammation for some time before there is any evidence of joint damage, but once joint damage begins the risk of continued damage is substantially greater.

Phase I may be viewed as ending upon the occurrence of a precipitating event which signals the beginning of a fundamentally different period (phase II) in which activity and damage are realized. The nature of the morbidity process will drive the specification of the stochastic model for this phase. Often the dynamics of the disease process are sufficiently distinct in this phase that it is natural to define the time origin as the time of the transition from phase I to phase II. With this in mind we formulate a partially semi-Markov two-phase model in which one part characterizes the duration of phase I and another part characterizes the dynamic disease process during the second phase. The term partially semi-Markov is used because the time origin is redefined only once, at the start of the second phase. This model can be used to separately examine prognostic factors for the length of the indolent phase as well as factors prognostic for the nature and rate of change in the active phase. In some settings this will offer a more appropriate representation of complex multi-phase disease processes, can help identify different types of risk factors, and could yield more accurate prediction models.

The remainder of the paper is organized as follows. In the next sub-section we describe the data from the University of Toronto Psoriatic Arthritis Cohort which motivates this work. In Section 2 we define notation and describe the two-part model using a general multistate process to characterize the second phase where the time origin for the second phase is the time of the precipitative event. We also discuss likelihood construction when individuals are examined intermittently rendering the time of the precipitating event and subsequent transition times as interval-censored. In Section 3 we consider a special case of the general phase II model of Section 2 which is specified to correspond to the data from the motivating study. Specifically, the response of interest is intermittent counts of the number of damaged joints experienced by patients with a rheumatological disease, so we consider an analysis based on proportional rate models. We then develop an expectation-maximization algorithm (Dempster et al., 1977) for estimation under a model with piecewise constant intensities. A computationally more convenient two-stage estimation procedure is discussed in Section 4 in which the parameters in the hazard for the end of phase I are estimated using standard likelihood for interval-censored data. The results of simulation studies examining the finite sample performance of estimators obtained by

maximum likelihood and the two-stage procedure are given in Section 5, along with an application to the motivating study. Concluding remarks and topics for further research are provided in Section 6. Derivations of the formula for large sample variance estimations under maximum likelihood and two-stage estimation are given in appendices.

1.2 The University of Toronto Psoriatic Arthritis Cohort

The Centre for Prognosis Studies in Rheumatic Disease is a tertiary care center at the Toronto Western Hospital which treats patients with a variety of rheumatological conditions and maintains several clinic registries with prospective follow-up. One registry is of patients with psoriatic arthritis (PsA), an immunological disease which features both skin (psoriasis) and joint (arthritis) involvement. The psoriatic aspect of the condition arises from an overproduction of new skin cells resulting in red and white scaly patches of skin frequently located on the elbows, knees and scalp. As with other arthritic conditions, this disease can result in considerable inflammation and ultimately destruction of joints, which can lead to serious disability and poor quality of life (Chandran et al., 2010). This registry was established in 1976 and has been recruiting and following patients since its inception, and today it is one of the largest cohorts of patients with PsA in the world.

Patients in this registry undergo a detailed clinical and radiological examination upon entry to the clinic, and provide serum samples for genetic testing. Follow-up clinical and radiological assessments are scheduled annually and biannually respectively in order to track changes in joint damage. At each radiological assessment the degree of damage is recorded in sixty-four joints on a five-point scale (Rahman et al., 1998). To date 1191 patients have been recruited to the University of Toronto Psoriatic Arthritis Clinic. Of these 604 have undergone genetic testing to determine their human leukocyte antigen profile. Among these individuals the median time from clinic entry to the last radiological assessment is 6.3 years with a median of 3 radiological assessments per patient.

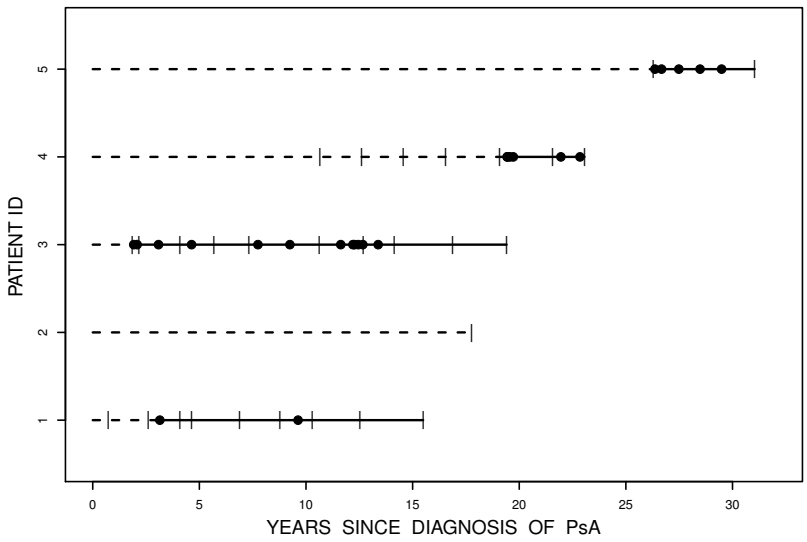

Figure 1: Plot of assessment times (hatch marks) and time of radiological damaged joints detected between assessments (solid points) from onset of psoriatic arthritis for a selected sample of patients from the University of Toronto Psoriatic Arthritis Clinic. The dashed line denotes time from disease onset to first occurrence of joint damage, and the solid line denotes the period of disease progression following onset of damage.

We focus our modeling here on the accumulation of joint damage reflected by the total number of joints with at least grade 4 damage according to the Steinbrocker scoring system (Wu and Cook,

2015). Figure 1 shows the time course of damage for a sample of five individuals. The horizontal axis reflects the time from disease onset and the length of the individual lines reflects the extent of follow-up of each individual; visits at which joint counts are made are represented by vertical tick marks. The dashed portion of each individuals timeline reflects the period during which no joint damage occurred, and the solid parts of the timelines reflect periods over which joint damage occurred. Of course the precise times at which joints became damaged are not available so for graphical illustration times were uniformly generated over the intervals during which they were known to occur; the dots are located at the resulting times.

From these illustrative timelines is apparent that some individuals develop joint damage shortly after diagnosis (e.g. individuals 1 and 3) while some enjoy a long period of time without damage (e.g. individuals 2, 4 and 5). Moreover, it appears that once the first joint becomes damaged, for some individuals other joints rapidly become damaged while for other individuals the rate of subsequent damage is low. Separate modeling of these two aspects of the disease process (the duration of the indolent phase, and the rate of damage in the active phase) are the focus of the two-part model we describe in the next section.

2 Model Formulation and Likelihood under Intermittent Observation

2.1 General Formulation of a Two-Phase Model

We consider chronic diseases that feature a variable and potentially long phase I during which there are no clinically important manifestations of disease in affected individuals. If t denotes the time since disease onset, we let T_1 be a random variable representing the duration of phase I with t_1 representing its realization. The second phase of the process takes place over $t > t_1$ when the disease is in an active period manifest by, for example, the occurrence of exacerbations or flares of symptoms, disability, or in the motivating rheumatological context, joint damage and destruction. The variable duration of the indolent period (phase I) and the distinct nature of disease activity in the active period (phase II) motivates the time scale for phase II defined based on $t^* = t - t_1$, which is simply the time since the end of phase I.

For a general presentation we consider a multistate disease process $\{Z(t^*), 0 < t^*\}$ for phase II which has a countable state space with states labeled by positive integers $\{1, 2, \ldots, \}$. Transitions may occur between any states in this general formulation with transitions governed, given $T_1 = t_1$, by Markov intensities in terms of time t^*. In Section 3.1 we consider the special case of a progressive multistate model for which only $k - 1 \to k$ transitions can occur.

To unify the notation for the two phases we augment the state space for the process in phase II to include a state 0 representing the status prior to T_1, and write $\bar{Z}(t) = Z(t^*)I(t_1 \leq t)$ and consider models for the stochastic process $\left\{\bar{Z}(t), 0 < t\right\}$; note that $Z(t^*) = \bar{Z}(t_1 + t^*)$. We let X be a $p \times 1$ vector of fixed covariates. The two-phase model can be defined by first considering the hazard for the end of phase I, defined by

$$\lim_{\Delta t \downarrow 0} \frac{P(t \leq T_1 < t + \Delta t | t \leq T_1, X)}{\Delta t} = h(t|X) \ . \tag{2.1}$$

Covariate effects can be modeled using proportional (Cox, 1972), additive (Aalen, 1989), or hybrid Cox-Aalen models (Martinussen and Scheike, 2007). We let $\mathcal{H}(t^*) = \{Z(u), 0 < u < t^*, x\}$ be the history of the process in phase II, and dynamic aspects of the process can be modeled through intensity functions (Andersen et al., 2012) given by

$$\lim_{\Delta t \downarrow 0} \frac{P\left(Z\left(t^* + \Delta t^-\right) = k | Z(t^{*-}) = j, \mathcal{H}(t^*)\right)}{\Delta t} = Y_j(t^*)\lambda_{jk}(t^*|\mathcal{H}(t^*)) \ , \tag{2.2}$$

where $Y_j(t^*) = I(Z(t^{*-}) = j)$, $j \in \{1, 2, \ldots\}$. If $\bar{\mathcal{H}}(t) = \left\{\bar{Z}(u), 0 < u < t, x\right\}$ denotes the history since the time of disease onset, then the intensity

$$\lim_{\Delta t \downarrow 0} \frac{P\left(\bar{Z}(t + \Delta t^-) = k | \bar{Z}(t^-) = j, \bar{\mathcal{H}}(t)\right)}{\Delta t} = \bar{Y}_j(t)\bar{\lambda}_{jk}(t|\bar{\mathcal{H}}(t)) \,, \tag{2.3}$$

governs the full process from disease onset, where $\bar{Y}_j(t) = I(\bar{Z}(t^-) = j)$ indicates whether an individual is at risk of a transition out of state $j \in \{0, 1, \ldots\}$ at time t. Note that if we denote (2.1) as $\lambda_{01}(t|\bar{\mathcal{H}}(t)\}$, then we can write $\bar{\lambda}_{jk}(t|\bar{\mathcal{H}}(t)) = \lambda_{jk}(B(t)|\bar{\mathcal{H}}(t))$ where $B(t) = I(t \leq t_1)t + I(t_1 < t)(t - t_1)$. Thus the process $\{\bar{Z}(t), 0 < t\}$ has a countable number of states in the state space and a semi-Markov feature in that the relevant time scale for the second phase of the disease process is the time since the end of phase I. With this time scale the process in phase II is Markov, but we refer to the process as a whole as partially semi-Markov process.

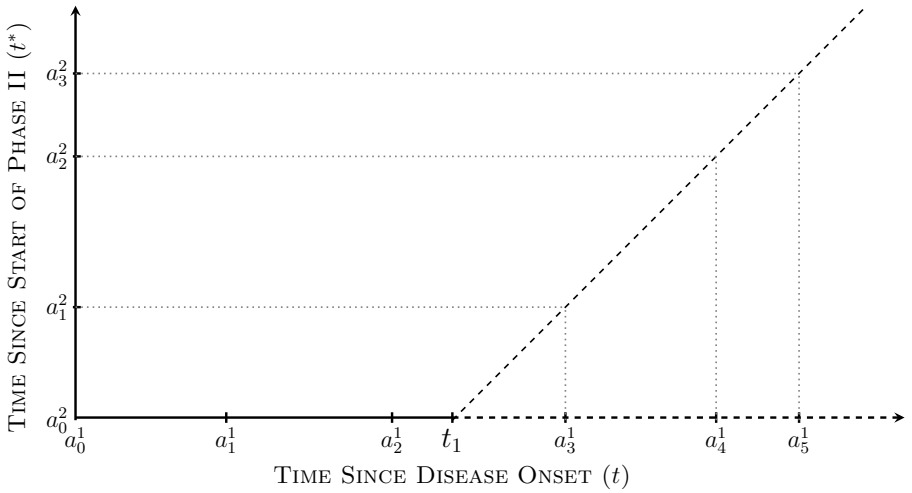

Figure 2: Lexis diagram of event and assessment times on the scale of disease duration (t) on the horizontal axis and the time since start of phase II (t^*) on the vertical axis.

The probability of a particular path $\mathcal{P}$ of this multistate process given $X = x$ is

$$\prod_{j=0}^{\infty} \prod_{k \in \mathcal{Z}_j} \left[\left\{ \prod_{t_r \in \bar{\mathcal{D}}_{jk}} \bar{\lambda}_{jk}(t_r|\bar{\mathcal{H}}(t_r)) \right\} \exp\left(- \int_0^{\infty} \bar{Y}_j(u)\bar{\lambda}_{jk}(u|\bar{\mathcal{H}}(u))\mathrm{d}u \right) \right] \,, \tag{2.4}$$

where $\mathcal{Z}_j$ is the set of states that can be entered directly from state j and $\bar{\mathcal{D}}_{jk}$ is the set of $j \to k$ transition times (Andersen and Keiding, 2002). This can be written more explicitly as

$$\begin{aligned} &\lambda_{01}(t_1|x) \exp\left(- \int_0^{\infty} \bar{Y}_0(u)\lambda_{01}(u|x)\mathrm{d}u \right) \\ &\times \left[\prod_{j=1}^{\infty} \prod_{k \in \mathcal{Z}_k} \left\{ \prod_{t_r^* \in \mathcal{D}_{jk}} \lambda_{jk}(t_r^*|\mathcal{H}(t_r^*)) \right\} \exp\left(- \int_0^{\infty} Y_j(u)\lambda_{jk}(u|\mathcal{H}(u))\mathrm{d}u \right) \right] \,, \end{aligned} \tag{2.5}$$

where if $t_r \in \bar{\mathcal{D}}_{jk}$, each $t_r^* \in \mathcal{D}_{jk}$ can be expressible as $t_r^* = t_r - t_1$. A slightly modified version of this probability expression can be derived for likelihood contributions when processes are under conditionally independent and non-informative censoring. Instead of pursuing this we consider next the problem of estimation and inference when such processes are under intermittent observation so that all event times are interval-censored.

2.2 Intermittent Assessment and Interval-Censored Data

Here we consider the setting in which individuals are intermittently assessed and discuss the likelihood construction; for simplicity in what follows, we consider the contribution from a single individual. Let $a_0 = 0$ denote the onset time of disease and $a_1 < \cdots < a_R$ denote the times of the R assessments at which point the individual's condition, and hence response status, is determined. The observed history at a_r^- is denoted by $H(a_r) = \left\{(a_\ell, \bar{Z}(a_\ell)), \ell = 0, 1, \ldots, r-1, X\right\}$, where we use a standard font for $H(\cdot)$ to distinguish it from the history of the process in continuous time. With fixed covariates, the full likelihood is

$$L \propto P\left(\bar{Z}(a_0), A_0 = a_0, X\right) \times \prod_{r=1}^{R} P\left(\bar{Z}(a_r), A_r = a_r | H(a_r)\right) \ . \tag{2.6}$$

We can omit the first term in the full likelihood if we condition on the covariate and the state occupied (0) at the onset of disease. We also assume the "sequential missing at random" condition (Hogan et al., 2004) holds so that if an individual is observed up to a_{r-1}, then conditional on the event history at that time, the probability they are lost to follow-up and not observed at a_r cannot depend on events in $[a_{r-1}, a_r)$. We also assume the event process and inspection process are conditionally independent and that the inspection process is non-informative. Under these assumptions, we can focus on the partial likelihood of the form

$$L \propto \prod_{r=1}^{R} P\left(\bar{Z}(a_r) | H(a_r)\right) \ . \tag{2.7}$$

This observed data partial likelihood (2.7) can be maximized directly, but this can be challenging if the dimension of parameters is high and the expression of this likelihood is complicated due to intermittent assessment. Therefore, an expectation-maximization (EM) algorithm (Dempster et al., 1977) can alternatively be used with a complete data likelihood analogous to observed data likelihood where missing variables, in this case the transition time from phase I to phase II, are part of the complete data. This is a particularly attractive approach for the setting of piecewise constant intensities which we consider in the next section.

3 An Expectation-Maximization Algorithm

3.1 The Complete Data Log-likelihood

The complete data likelihood (2.5) is given in general form for the case in which we consider the event times as observed, or subject at most to right-censoring. In this section, we redefine the notation by giving a superscript 1 or 2 to denote the part. Here we consider the setting with interval-censored data in phase I and let $a_0^1 = 0$ denote the onset of disease and $a_1^1 < \cdots < a_{R_1}^1$ denote the times of R_1 assessments at which point the individual's disease stage is determined. For information in phase II it is helpful to let $a_0^2 = 0$ denote the start time of phase II and $a_1^2 < \cdots < a_{R_2}^2$ denote the times of the R_2 radiological assessments during phase II. With the process in phase II a recurrent event process we can also let $n_r = \bar{Z}(a_r^2) - \bar{Z}(a_{r-1}^2)$ denote the number of events over the interval $\mathcal{A}_r = (a_{r-1}^2, a_r^2]$, $r = 1, \ldots, R_2$.

We adopt a Poisson process model for phase II such that $\lambda_{k,k+1}(t^* | \mathcal{H}(t^*)) = \rho(t^* | x)$ and write the complete data likelihood (2.5) as

$$L \propto \lambda_{01}(t_1|X) \exp\left(-\int_0^\infty \bar{Y}_0(u)\lambda_{01}(u|X)\mathrm{d}u\right) \times \prod_{r=1}^{R_2} \left[\frac{1}{n_r!}\left\{\int_{a_{r-1}^2}^{a_r^2} \rho(u|X)\mathrm{d}u\right\}^{n_r} \exp\left\{-\int_{a_{r-1}^2}^{a_r^2} \rho(u|X)\mathrm{d}u\right\}\right] \ . \tag{3.1}$$

We consider multiplicative models of the form $\lambda_{01}(t_1|X;\theta_1) = h_0(t_1;\alpha_1)\exp(X'\beta_1)$ for $T_1|X$ and $\lambda_{k,k+1}(t^*|\mathcal{H}(t^*)) = \rho_0(t^*;\alpha_2)\exp(X'\beta_2)$ for the recurrent event process in phase II, where $k \geq 1$, where α_1 indexes the baseline hazard function, α_2 indexes the baseline rate function, $\theta_1 = (\alpha_1', \beta_1')'$, $\theta_2 = (\alpha_2', \beta_2')'$ and $\theta = (\theta_1', \theta_2')'$. A weakly parametric piecewise exponential baseline hazard is adopted for the duration of phase I and a piecewise constant baseline rate model is adopted for the recurrent event process during phase II. These require specification of break-points where the baseline hazard and rate functions can take on different values and we denote these by $0 = b_0^1 < b_1^1 < \cdots < b_{K_1}^1$ and $0 = b_0^2 < b_1^2 < \cdots < b_{K_2}^2$ respectively. Then we let

$$\begin{aligned} h_0(t;\alpha_1) &= \alpha_{1k} \quad \text{if } t \in \mathcal{B}_k^1 = [b_{k-1}^1, b_k^1) \quad k = 1, \ldots, K_1, \\ \rho_0(t^*;\alpha_2) &= \alpha_{2k} \quad \text{if } t^* \in \mathcal{B}_k^2 = [b_{k-1}^2, b_k^2) \quad k = 1, \ldots, K_2 \ , \end{aligned} \tag{3.2}$$

respectively. We consider all the subintervals $\mathcal{C}_{rk} = \mathcal{A}_r \cap \mathcal{B}_k^2$ of length u_{rk} and let n_{rk} denote the unobserved number of events over $\mathcal{C}_{rk}$ such that $\sum_{k=1}^{K_2} n_{rk} = n_r$, $r = 1, \ldots, R_2$. Since $N_{rk}|T_1, X \sim$ Poisson(μ_{rk}) , where $\mu_{rk} = \alpha_{2k} u_{rk} \exp(X'\beta_2)$, then

$$E\left(N_{rk}|T_1, X, N_r\right) = n_r \cdot \alpha_{2k} u_{rk} / \textstyle\sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \ .$$

The complete data log-likelihood is then

$$\log L_{\mathrm{C}}(\theta) = \log L_{\mathrm{C1}}(\theta_1) + \log L_{\mathrm{C2}}(\theta_2) \ , \tag{3.3}$$

where

$$\log L_{\mathrm{C1}}(\theta_1) = \delta_1 \left\{ \sum_{k=1}^{K_1} I_k(t_1) \left(\log \alpha_{1k} + X'\beta_1\right) - \sum_{k=1}^{K_1} \alpha_{1k} W_k(t_1) e^{X'\beta_1} \right\} - (1-\delta_1) \sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{X'\beta_1} \ , \tag{3.4}$$

$$\log L_{\mathrm{C2}}(\theta_2) = \delta_1 \left\{ \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} n_{rk} \left(\log \alpha_{2k} + X'\beta_2\right) - \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} e^{X'\beta_2} \right\} \ , \tag{3.5}$$

$I_k(u) = I(u \in \mathcal{B}_k^1)$ and $W_k(u) = \int_0^u I_k(s)\mathrm{d}s$ is the total time at risk in $\mathcal{B}_k^1$ over the interval $(0, u]$, $k = 1, \ldots, K_1$ and $\delta_1 = I(T_1 < a_{R_1})$ is the status indicator whether the precipitating event is observed.

3.2 The EM Algorithm for Maximum Likelihood Estimation

At the vth iteration of the expectation-maximization (EM) algorithm, the E-Step is to take the conditional expectation

$$Q(\theta;\theta^{(v)}) = Q_1(\theta_1;\theta^{(v)}) + Q_2(\theta_2;\theta^{(v)}) \ , \tag{3.6}$$

where $Q_1(\theta_1;\theta^{(v)}) = E\left[\log L_{\mathrm{C1}}(\theta_1)|D;\theta^{(v)}\right]$ and $Q_2(\theta_2;\theta^{(v)}) = E\left[\log L_{\mathrm{C2}}(\theta_2)|D;\theta^{(v)}\right]$, where the observed data is $D = \left\{(a_r, \bar{Z}(a_r)), r = 0, 1, \ldots, R_1, X\right\}$. The unobserved quantities in the complete data log likelihood $I_k(t_1)$, $W_k(t_1)$, n_{rk} and u_{rk} are all functions of T_1. Thus, their conditional expectations given the current estimates of parameters and the observed data D can be evaluated through

$$f_{t_1|D}(t_1|D;\theta) = \frac{f_1(t_1) \times \prod_{r=1}^{R_2} f_2(n_r|t_1)}{\int_{L_1}^{R_1} f_1(u_1) \times \prod_{r=1}^{R_2} f_2(n_r|u_1)\ \mathrm{d}u_1} \ , \tag{3.7}$$

where

$$f(t_1|X) = \prod_{k=1}^{K_1} \left\{ [\alpha_{1k} \exp(X'\beta_1)]^{I_k(t_1)} \cdot \exp\left(-\alpha_{1k} W_k(t_1) \exp(X'\beta_1)\right) \right\}$$

and

$$f_2(n_r|t_1, X) = \left[\sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \exp(X'\beta_2) \right]^{n_r} \cdot \exp\left(- \sum_{k=1}^{K_2} \alpha_{2k} u_{rk} \exp(X'\beta_2) \right) .$$

The M-Step involves maximizing $Q(\theta; \theta^{(v)})$ with respect to θ and get the updated estimate $\theta^{(v+1)}$. By reparametrization, we can write $Q(\theta; \theta^{(v)})$ in the form of a Poisson log-likelihood and use existing software for generalized linear model to maximize following the creation of a pseudo-dataset. We iterate between the E-step and M-step until the convergence criterion $\left| \left(\theta^{(v+1)} - \theta^{(v)}\right) / \theta^{(v)} \right| < \epsilon$ is achieved where ϵ is the user-specified tolerance. The details of the EM algorithm and the calculation of conditional expectations are given in Appendix A.

4 Two-Stage Estimation

Instead of simultaneously estimating all the parameters in the full likelihood function (3.3), a two-stage estimation procedure can be adopted. Under this approach in the first stage we note that we can simply view T_1 as an interval-censored failure time with a hazard function indexed by θ_1. For this the pertinent data can be denoted by $\mathcal{C}_1 = [L_1, R_1)$, the interval known to contain T_1 and X. Here we let $Q_{\text{I}}(\cdot)$ denote the corresponding function in (3.6) under a two-stage procedure where

$$Q_{\text{I}}(\theta_1; \theta_1^{(v)}) = \delta_1 \left[\sum_{k=1}^{K_1} \hat{\iota}_k^{(v)} \left(\log \alpha_{1k} + X'\beta_1 \right) - \sum_{k=1}^{K_1} \widehat{\omega}_k^{(v)} \alpha_{1k} e^{X'\beta_1} \right] - (1 - \delta_1) \sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{X'\beta_1} , \tag{4.1}$$

$\hat{\iota}_k^{(v)} = E\{I_k(t_1)|\mathcal{C}_1, x; \theta_1^{(v)}\}$ and $\widehat{\omega}_k^{(v)} = E\{W_k(t_1)|\mathcal{C}_1, x; \theta_1^{(v)}\}$. The conditional distribution of T_1 given $\mathcal{C}_1$ and X takes on a simpler form in this framework with $f(t_1|\mathcal{C}_1, X; \theta_1^{(v)}) = f_1(t_1) / \int_{L_1}^{R_1} f_1(u_1) du_1$ given by

$$f(t_1|\mathcal{C}_1, X; \theta_1^{(v)}) = \frac{[\prod_{k=1}^{K_1} \alpha_{1k}^{I_k(t_1)}] \times \exp(-\sum_{k=1}^{K_1} \alpha_{1k} W_k(t_1) \exp(X'\beta_1))}{\int_{L_1}^{R_1} [\prod_{k=1}^{K_1} \alpha_{1k}^{I_k(u_1)}] \times \exp(-\sum_{k=1}^{K_1} \alpha_{1k} W_k(u_1) \exp(X'\beta_1)) du_1} . \tag{4.2}$$

The expectations are therefore easier to carry out, and the maximization step is as before. Specifically (4.1) can be written as a Poisson log-likelihood and existing software can be used to maximize it following the creation of a pseudo-dataset as in Section 3.

In the second stage, θ_2 can be estimated via a modified expectation-maximization algorithm defined by plugging in the estimates of $\widehat{\theta}_1$ from stage one into the function $Q_{\text{II}}(\theta_2; \widehat{\theta}_1, \theta_2^{(v)})$ defined as

$$Q_{\text{II}}(\theta_2; \widehat{\theta}_1, \theta_2^{(v)}) = \sum_{r=1}^{R_2} \sum_{k=1}^{K_2} \delta_1 \left\{ \widehat{n}_{rk}^{(v)} \left(\log \alpha_{2k} + X'\beta_2 \right) - \alpha_{2k} \widehat{u}_{rk}^{(v)} \exp(X'\beta_2) \right\} , \tag{4.3}$$

where $\widehat{u}_{rk} = E[u_{rk}|D; \widehat{\theta}_1, \theta_2^{(v)}]$ and $\widehat{n}_{rk}^{(v)} = E[n_r \alpha_{2k}^{(v)} u_{rk} / \sum_{k=1}^{K_2} \alpha_{2k}^{(v)} u_{rk} | D; \widehat{\theta}_1, \theta_2^{(v)}]$. The conditional expectations in the second stage are based on (3.7) evaluated at $\widehat{\theta}_1$ and $\theta_2^{(v)}$ given the full set of observed data D. The objective function (4.3) can be rewritten to take the form of a Poisson log-likelihood and maximized using existing software as before. This two-stage estimation approach is quite similar to the method of simultaneous estimation we described in Section 3.2; however, this approach is computationally faster, particularly when the number of parameters is large. Variance estimation of the resulting estimation is discussed in Appendix C. We comment further on the potential uses of this two-stage procedure in the Discussion.

5 Simulation Studies and Application

5.1 Design and Interpretation of Simulation Studies

In this section, a simulation study is conducted to demonstrate the performance of proposed two-phase model. For each individual i, a $p \times 1$ covariate vector X_i is generated from a multivariate normal distribution with mean 0 and a covariance matrix Σ, where $p = 2$, $\Sigma_{ij} = \varrho^{|i-j|}$ and $\varrho = 0.5$. The duration of the indolent phase T_1 is generated from an exponential distribution with rate $\alpha_1 \exp(X_i'\beta_1)$. We set $\beta_1 = (0.5, 0.5)'$ and solve for the value of the baseline rate α_1 such that $F(C) = \int_x P(T_1 < C|x)P(x)dx = 0.8$, where $C = 50$ is the administrative censoring time. The gap times between the consecutive events are generated by an exponential distribution with rate $\alpha_2 \exp(X_i'\beta_2)$, where $\beta_2 = (-0.5, -0.5)$ and $\alpha_2 = 0.5$.

We let R_i denote the number of assessments for individual i, which is generated according to a truncated Poisson distribution to ensure at least one follow-up assessment, with

$$P(R_i = r_i | R_i \geq 1; \mu) = \frac{\mu^{r_i} \exp(-\mu)}{r_i! \{1 - \exp(-\mu)\}}, r_i = 1, \ldots .$$

where $\mu = 10$. The R_i inspection times $0 < a_{i1} < \cdots < a_{iR_i} < C$ are then uniformly distributed over $[0, C]$. The number of events occurring between assessments are then $m_{ir} = \sum_{j=1}^{n_i} I(a_{i,r-1} < t_{ij} \leq a_{i,r})$, $r = 1, \ldots, R_i$. We consider a sample size of $m = 500$ and simulate five hundred datasets ($nsim = 500$). For each dataset, we fit the proposed two-phase model by the EM algorithm under both simultaneous (maximum likelihood) and two-stage estimation; the empirical performance of estimators are shown in Table 1. The break-points were chosen for the phase I model as the quartiles of the baseline survival distribution. For the second phase, they are chosen to be equally spaced over $[0, C - Q_1^{50}]$ (i.e. we used $(C - Q_1^{50})/4$, $(C - Q_1^{50})/2$ and $3(C - Q_1^{50})/4$) where Q_1^{50} is the median of T_1; in Table 1 for example $(C - Q_1^{50})/4 = 9.23$. Standard errors for the maximum likelihood estimators were obtained by Louis (Louis, 1982) (see Appendix B) and using estimating function theory (see Appendix C). The empirical coverage probabilities were computed as the proportion of all simulated datasets for which the 95% confidence interval contained the true parameter value.

The empirical performance of the estimators using both estimation approaches are shown in Table 1, where the empirical biases (EBIAS) are generally small. There is good agreement between the empirical standard errors (ESE) and average standard errors (ASE) obtained by Louis (Louis, 1982) or the methods of Appendix C respectively and the empirical coverage probabilities (ECP) are all compatible with the nominal level. From the simulation results, we can conclude that both estimation approaches give good performance; the empirical biases are relatively small and the empirical coverage probabilities are all compatible with the nominal 95% level. Moreover, there is relatively little efficiency loss from the two-stage estimation procedure.

5.2 Application of Psoriatic Arthritis Data

Here we consider the data on joint damage in patients with psoriatic arthritis from the University of Toronto Psoriatic Arthritis Clinic. Specific interest lies in examining the effects of human leukocyte antigen (HLA) markers on the duration of the indolent phase following diagnosis and on the rate of joint damage following the end of the indolent phase. The breakpoints for the model of the duration of the indolent phase are 3.5, 9.2, 13.7 and 26 years, specified to correspond to points yielding roughly equal increments in the nonparametric Turnbull estimate of the the cumulative probability function for the time to the precipitating event (the time the first joint becomes damaged). The breakpoints for the second phase were taken as 8.2, 12.6, 17.0 and 23.5 years, likewise corresponding to roughly equal increments in the estimate of the mean function for the cumulative number of damaged joints

Table 1: Empirical performance† of estimators; sample size $m = 500$, number of simulations $nsim = 500$, $\alpha_1 = 0.036$, $\alpha_2 = 0.5$, $\beta_1 = (0.5, 0.5)$, $\beta_2 = (-0.5, -0.5)$; ASE are average of standard errors estimated via methods in Appendix B (Maximum Likelihood) and Appendix C (Two-Stage Estimation).

		MAXIMUM LIKELIHOOD				TWO-STAGE ESTIMATION			
PIECE	PARAMETER	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
		PHASE I: ONSET OF DAMAGE							
$[0.00, 5.38)$	α_{11}	-0.014	0.447	0.431	94.6	-0.005	0.475	0.457	93.8
$[5.38, 13.05)$	α_{12}	0.008	0.445	0.452	95.6	0.004	0.484	0.496	95.4
$[13.05, 25.26)$	α_{13}	0.021	0.413	0.427	95.6	0.008	0.447	0.458	96.4
$[25.26, 50.00)$	α_{14}	0.042	0.450	0.440	94.4	0.052	0.456	0.451	94.4
	β_{11}	0.968	6.491	6.494	95.4	1.000	6.449	6.538	95.6
	β_{12}	0.654	6.431	6.447	96.6	0.621	6.467	6.488	97.2
		PHASE II: PROGRESSION OF DAMAGE							
$[0.00, 9.23)$	α_{21}	-0.163	1.626	1.622	95.2	-0.167	1.626	1.623	95.0
$[9.23, 18.46)$	α_{22}	0.031	1.763	1.815	95.6	0.030	1.760	1.815	95.6
$[18.46, 27.68)$	α_{23}	0.104	2.071	1.994	94.8	0.103	2.068	1.994	94.8
$[27.68, 50.00)$	α_{24}	-0.003	1.781	1.702	93.2	-0.003	1.780	1.702	93.2
	β_{21}	-0.117	1.669	1.691	95.4	-0.118	1.669	1.691	95.2
	β_{22}	0.039	1.617	1.686	96.4	0.037	1.617	1.686	96.4

† EBIAS, ESE and ASE reported are $\times 10^2$

$N(t)$ obtained based on isotonic regression. We examine the effects of HLA markers selected based on the results of Wu and Cook (Wu and Cook, 2015), while controlling for gender and patient age.

As in the empirical studies, we find from the results in Table 2 that there is good agreement in the estimates obtained by the simultaneous and two-stage estimation procedures. We therefore discuss the results of maximum likelihood estimation here. Among the HLA markers, HLA-A11, HLA-A25, HLA-A29, HLA-A30, HLA-C03 and HA-DRB1-10 had insignificant association with the duration of the indolent phase but their presence was associated with a significant reduction of the rate of damage in the active phase of the disease. For HLA-A11 for example, the relative rate of damage in the active phase associated with the presence of HLA-A11 is $RR = 0.70$ (95% CI : 0.53, 0.91; $p = 0.0087$); the corresponding relative rates for the other markers were HLA-A25 $RR = 0.07$ (95% CI : 0.01, 0.53; $p = 0.0096$), HLA-A29 $RR = 0.25$ (95% CI : 0.15, 0.43; $p < 0.0001$), HLA-A30 $RR = 0.78$ (95% CI : 0.63, 0.97; $p = 0.0235$), HLA-C03 $RR = 0.57$ (95% CI : 0.47, 0.70; $p < 0.0001$), and HLA-DRB1-10 $RR = 0.04$ (95% CI : 0.01, 0.27; $p = 0.0011$). Moreover, there is significant evidence that the effect of HLA-A25, HLA-A30, HLA-C03, and HLA-DRB1-10 on the duration of the indolent phase and on damage progression are different; see the last column of Table 2 for the homogeneity $p-$values. HLA-B27 is a known risk factor for disease progression in PsA and here we find its presence is associated with both a shorter indolent phase and more rapid disease progression; the same can be said for HLA-DQB1-02.

Figure 3 displays estimates of the probability of having at least one damaged joint as measured by the time from disease onset (left panel) as well as the expected number of damaged joints from the time of disease onset (right panel). The dashed lines in the left panel represents a nonparametric Turn-

Table 2: Results of fitting a piecewise constant baseline hazard model for the duration of the indolent period and a piecewise constant baseline rate model for the occurrence of joint damage under simultaneous (ML) and two-stage estimation; $p-$values are based on Wald tests.

		MAXIMUM LIKELIHOOD				TWO-STAGE ESTIMATION			
HLA Marker	PHASE	EST	SE	p	$p^\dagger$	EST	SE	p	$p^\dagger$
HLA-A11	I	-0.280	0.274	0.3079		-0.315	0.274	0.2506	
	II	-0.363	0.138	0.0087	**0.7876**	-0.358	0.139	0.0098	**0.8905**
HLA-A25	I	-0.211	0.597	0.7242		-0.159	0.597	0.7900	
	II	-2.627	1.014	0.0096	**0.0407**	-2.614	1.015	0.0100	**0.0378**
HLA-A29	I	-0.597	0.371	0.1076		-0.601	0.371	0.1055	
	II	-1.377	0.270	< 0.0001	**0.0899**	-1.375	0.270	< 0.0001	**0.0925**
HLA-A30	I	0.458	0.295	0.1208		0.364	0.299	0.2243	
	II	-0.250	0.110	0.0235	**0.0256**	-0.249	0.110	0.0244	**0.0564**
HLA-B27	I	0.468	0.183	0.0105		0.490	0.183	0.0074	
	II	0.235	0.067	0.0004	**0.2333**	0.237	0.067	0.0004	**0.1957**
HLA-C03	I	0.014	0.219	0.9480		0.017	0.220	0.9367	
	II	-0.563	0.103	< 0.0001	**0.0178**	-0.566	0.103	< 0.0001	**0.0169**
HLA-C04	I	-0.012	0.224	0.9576		0.011	0.225	0.9606	
	II	-0.120	0.106	0.2565	**0.6650**	-0.122	0.107	0.2568	**0.5985**
HLA-DQB1-02	I	0.386	0.164	0.0187		0.394	0.164	0.0163	
	II	0.249	0.062	< 0.0001	**0.4365**	0.252	0.063	< 0.0001	**0.4215**
HLA-DRB1-10	I	0.203	0.594	0.7326		0.195	0.594	0.7428	
	II	-3.280	1.002	0.0011	**0.0028**	-3.275	1.002	0.0011	**0.0029**

$p^\dagger$: a p-value from a test of homogeneity.

bull estimate (Turnbull, 1976) of the distribution of the duration of phase I based on interval-censored T_1 times. The solid line denotes the piecewise constant estimate from our proposed two-phase model under simultaneous estimation of the phase I and II parameters. The dotted line gives an estimate obtained by fitting a Poisson process to the interval-grouped joint damage data (Lawless and Zhan, 1998) and computing the probability of no damaged joints. The close alignment of the nonparametric and proposed estimator suggests the piecewise constant hazards model gives a reasonable representation of the data and the large disparity between the Poisson-based estimate highlights the importance of accommodating the two phases of the disease process. The steep rise in the probability of at least one joint becoming damaged under the Poisson model is in conflict with the data.

The right panel of Figure 3 gives estimates of the expected number of damaged joints over time. A nonparametric isotonic estimate (Barlow et al., 1972) is displayed with a dashed line and a solid line gives the estimate under the proposed model with a piecewise constant baseline rate. Again Poisson model is used to fit the data as a benchmark and displayed with a dotted line. It can be seen that the benchmark analysis will overestimate the expected number of damaged joints and the proposed two-phase model is comparable with the nonparametric analyses. This demonstrates the proposed two-phase model fits the data well and supports the need to address the heterogeneity in the disease process in this way.

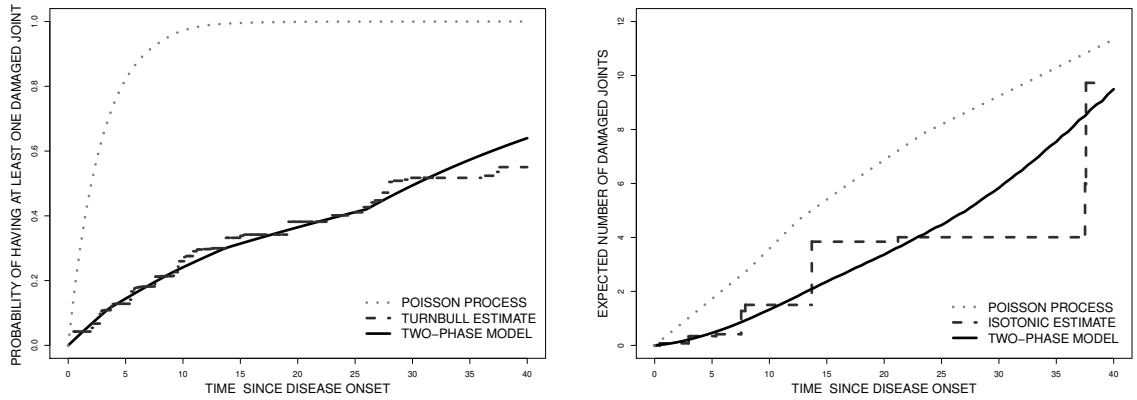

Figure 3: Estimates of the probability of having at least one damaged joint as measured from the time since disease onset (left panel) and the expected number of damaged joints as function of the time since disease onset (right panel); a benchmark analysis is based on a Poisson process with interval-grouped recurrent event data (Lawless and Zhan, 1998); the nonparametric estimate for the duration of phase I is based on a Turnbull estimate (Turnbull, 1976) and an isotonic estimate (Barlow et al., 1972) for the expected number of damaged joints respectively.

6 DISCUSSION

There are several avenues for generalizations of this work including the use of semiparametric models for the time from disease onset to the event signalling the beginning of the second phase of the process. Much work has been done in the last twenty years on the development of flexible regression methodology and statistical theory for the analysis of interval-censored failure time data (Sun, 2006). A more challenging generalization would be to relax the Markov assumption for the second phase process; this is challenging since not many alternative models can be easily fitted when processes are under intermittent observation. In the general multistate formulation of Section 2 much work has been done on methods for fitting and assessing Markov models in this setting but semi-Markov and models with hybrid time scales have seen little development. Mixed effect models which are Markov conditional on latent random effects however, have been developed and render more elaborate dependencies on the process history. The conditional Markov property enables one to fit these models even when the historical information is unobserved. For the recurrent event model we consider in Section 3 this would correspond to a mixed Poisson model for the second phase of the process which would be negative binomial if a gamma distributed random effect were introduced. Large data sets would be required to estimate parameters in this more flexible model.

Another generalization of interest would be to integrate survival data into the disease process. This would introduce an absorbing state which would of course terminate the disease process. The primary purpose of this method is to separate the two phases of the morbidity process and to obtain separate estimates of the effects of genetic markers; the effects of these markers are thought to be fairly robust to violations of the assumption of independent mortality but this represents an area worthy of development. Scientists at the University of Toronto Psoriatic Arthritis Cohort are undertaking tracing studies to collect data on survival status of individuals who have not been to the clinic for some time. With this additional information more comprehensive models could be considered which incorporate survival data.

We have carried out tests of the null hypothesis of common coefficients for the phase I and II regression models. If the null hypothesis is not rejected for a particular marker, one could consider fitting a model with the constraint that the effects are the same. The algorithm can be adapted to handle this but given the quite different interpretation of the effects in the two phases we have not considered that here. It would also be of interest to assess whether there is evidence of a need for the two-phase

model because estimation, inferences and model interpretation would be so much easier if the second phase model were adequate. Such a test would be analogous to tests for the need to accommodate a non-susceptible fraction in cure rate models, but in this context this is more challenging since the time scale for the second phase model is defined as the time from the precipitating event.

Identification of important genetic and soluble biomarkers is of primary interest in psoriatic arthritis and the two-phase model offers an important opportunity to identify factors that may be prognostic for different aspects of the disease process. Given that a particular marker may be entertained in both parts of the model one could consider the use of the group LASSO (Yuan and Lin, 2005, Wang and Leng, 2008) by defining pairs of coefficients for each marker, with one coefficient defined in the regression model for the phase I duration and another defined in the phase II model. Variations of this such as the sparse group LASSO (Simon et al., 2013) could be useful in GWAS analyses but the standard group LASSO would be sufficient for the analysis of haplotype data.

Often cohort data are created from registries which required individuals to have experienced some disease manifestation for enrolment. This can lead to a biased sampling scheme arising due to truncation of the disease process. Researchers may require individuals to not have experienced disease activity or damage to be eligible for an inception cohort, which would result in right-truncated interval-censored duration times for the first phase. Cohorts of individuals with advanced disease may require progression to some advanced state of the second phase process yielding right-truncated phase I and II data. The expectation-maximization algorithm we describe can be adapted to accommodate left-, right- and interval-truncation by the conceptualization of "ghosts" in the spirit of Turnbull (Turnbull, 1976). Such a complete data likelihood will be possible to fit with penalty terms using standard software for penalized Poisson regression.

APPENDIX

A EVALUATION OF $Q(\theta;\theta^{(v)})$ FOR MAXIMUM LIKELIHOOD ESTIMATION

A.1 DETAILS OF EM ALGORITHM

Here we show the details of the EM algorithm we described in Section 3.2. At vth iteration, we proceed as follows:

1. Evaluate $\tilde{\iota}_k^{(v)} = E[I_k(t_1)|D;\theta^{(v)}]$ and $\widetilde{\omega}_k^{(v)} = E[W_k(t_1)|D;\theta^{(v)}]$. Then

$$Q_1(\theta_1;\theta^{(v)}) = \delta_1 \left\{ \sum_{k=1}^{K_1} \tilde{\iota}_k^{(v)} \left(\log \alpha_{1k} + x'\beta_1\right) - \sum_{k=1}^{K_1} \alpha_{1k} \widetilde{\omega}_k^{(v)} e^{x'\beta_1} \right\} - (1-\delta_1) \left(\sum_{k=1}^{K_1} \alpha_{1k} W_k(a_{R_1}) e^{x'\beta_1} \right) . \tag{A.1}$$

2. Maximize $Q_1(\theta_1;\theta^{(v)})$ to get the updated estimate of θ_1, $\theta_1^{(v+1)}$.
Let $Z_k = (Z_{k1},\ldots,Z_{kK_1})'$ denote the indicator function, where $Z_{k\ell} = I(k=\ell), \ell = 1,\ldots,K_1$.
Let $\alpha_k = \log \alpha_{1k}$, $k = 1,\ldots,K_1$ and $\alpha = (\alpha_1,\ldots,\alpha_{K_1})'$, then we can write

$$Q_1(\theta;\theta^{(v)}) = \sum_{k=1}^{K_1} \left[\delta_1 \left\{ \tilde{\iota}_k^{(v)} \left(z_k'\alpha + x'\beta_1\right) - \widetilde{\omega}_k^{(v)} e^{z_k'\alpha + x'\beta_1} \right\} - (1-\delta_1) W_k(a_{R_1}) e^{z_k'\alpha + x'\beta_1} \right] . \tag{A.2}$$

We note that (A.2) has a Poisson form of log likelihood function, then we can use existing software (`glm`) to maximize it by creating a pseudo-dataset in the following format, as shown in Table A.1.

Table A.1: Pseudo-dataframe for the maximization of Q_1

ID (i)	piece (k)	Z_{k1}	Z_{k2}	$\cdots$	Z_{kK_1}	X_1	$\cdots$	X_p	Response	Offset
					$\delta_1 = 1$					
i	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{\iota}_1^{(v)}$	$\log \widetilde{\omega}_1^{(v)}$
i	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{\iota}_2^{(v)}$	$\log \widetilde{\omega}_2^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	K_1	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{\iota}_{K_1}^{(v)}$	$\log \widetilde{\omega}_{K_1}^{(v)}$
					$\delta_1 = 0$					
i	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	0	$\log W_1(a_{R_1})$
i	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	0	$\log W_2(a_{R_1})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
i	K_1	0	0	$\cdots$	1	x_1	$\cdots$	x_p	0	$\log W_{K_1}(a_{R_1})$

3. Evaluate $\widetilde{u}_{rk}^{(v)} = E[u_{rk}|D;\theta_1^{(v+1)},\theta_2^{(v)}]$ and $\widetilde{n}_{rk}^{(v)} = n_r E[\alpha_{2k}^{(v)} u_{rk} / \sum_{k=1}^{K_2} \alpha_{2k}^{(v)} u_{rk}|D;\theta_1^{(v+1)},\theta_2^{(v)}]$. Then

$$
\begin{aligned}
&Q_2(\theta_2;\theta_1^{(v+1)},\theta_2^{(v)}) \\
=&E_{T_1}E_{N_{rk}|T_1}\left[\log L_{\text{C2}}(\theta_2)|D;\theta_1^{(v+1)},\theta_2^{(v)}\right] \\
=&\delta_1 E\left[\sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\left\{n_r \frac{\alpha_{2k}^{(v)}u_{rk}}{\sum_{k=1}^{K_2}\alpha_{2k}^{(v)}u_{rk}}\left(\log\alpha_{2k}+x'\beta_2\right)-\alpha_{2k}u_{rk}\exp(x'\beta_2)\right\}|D;\theta_1^{(v+1)},\theta_2^{(v)}\right] \\
=&\sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\delta_1\left\{\widetilde{n}_{rk}^{(v)}\left(\log\alpha_{2k}+x'\beta_2\right)-\alpha_{2k}\widetilde{u}_{rk}^{(v)}\exp(x'\beta_2)\right\}\ .
\end{aligned}
\tag{A.3}
$$

4. Maximize $Q_2(\theta;\theta_1^{(v+1)},\theta_2^{(v)})$ to get the updated estimate of θ_2, $\theta_2^{(v+1)}$. Let $Z_k = (Z_{k1},\ldots,Z_{kK_2})'$ denote the indicator function, where $Z_{k\ell} = I(k=\ell), \ell = 1,\ldots,K_2$. Let $\gamma_k = \log\alpha_{2k}$, $k = 1,\ldots,K$ and $\gamma = (\gamma_1,\ldots,\gamma_{K_1})'$, then we can write

$$
Q_2(\theta;\theta_1^{(v+1)},\theta_2^{(v)}) = \sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\delta_1\left\{\widetilde{n}_{rk}^{(v)}\left(z_k'\gamma+x'\beta_2\right)-\widetilde{u}_{rk}^{(v)}\exp(z_k'\gamma+x'\beta_2)\right\}\ . \tag{A.4}
$$

We note that (A.4) has a Poisson form of log likelihood function, then we can use existing software (`glm`) to maximize it by creating a pseudo dataset in the following format, as shown in Table A.2.

A.2 Evaluations of the Conditional Expectations in the E-Step

Since all the unobserved quantities are related to T_1, we need conditional expectations in the form of

$$
\int_{L1}^{R_1} f(x)\mathrm{d}x\ .
$$

Due to the complicated nature of the function $f(x)$, closed form expressions are not available so we use numerical integration. Here we describe the Gaussian Quadrature which we used in the analyses.

Table A.2: Pseudo-dataframe for the maximization of Q_2

ID (i)	assess (r)	piece (k)	Z_{k1}	Z_{k2}	$\cdots$	Z_{kK_2}	X_1	$\cdots$	X_p	Response	Offset
i	1	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{11}^{(v)}$	$\log \widetilde{u}_{11}^{(v)}$
i	1	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{12}^{(v)}$	$\log \widetilde{u}_{12}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	
i	1	K_2	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{n}_{1K_2}^{(v)}$	$\log \widetilde{u}_{1K_2}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$				
i	R_2	1	1	0	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,1}^{(v)}$	$\log \widetilde{u}_{R_2,1}^{(v)}$
i	R_2	2	0	1	$\cdots$	0	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,2}^{(v)}$	$\log \widetilde{u}_{R_2,2}^{(v)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	
i	R_2	K_2	0	0	$\cdots$	1	x_1	$\cdots$	x_p	$\widetilde{n}_{R_2,K_2}^{(v)}$	$\log \widetilde{u}_{R_2,K_2}^{(v)}$

First we can use a linear transformation to change this integration into a new integration on the interval $(-1, 1)$.
Let $x = \phi(y) = \{y(R1 - L1) + L_1 + R_1\}/2$, so

$$\int_{L1}^{R_1} f(x)\mathrm{d}x = \int_{-1}^{1} f(\phi(y))\phi'(y)\mathrm{d}y \ .$$

Then using Chebyshev quadrature (Golub and Welsch, 1969) of the 1st kind with the weight function $w(y) = 1/\sqrt{1-y^2}$, we approximate the integration by:

$$\begin{aligned}\int_{L1}^{R_1} f(x)\mathrm{d}x = \int_{-1}^{1} f(\phi(y))\phi'(y)\mathrm{d}y &= \int_{-1}^{1} w(y)\frac{f(\phi(y))\phi'(y)}{w(y)}\mathrm{d}y \\ &\doteq \int_{-1}^{1} w(y)g(y)\mathrm{d}y = \sum_{s=1}^{N} w_s g(y_s) \ ,\end{aligned}$$

where w_s and y_s are the weights and nodes that are picked based on weight function $w(y)$. Monte Carlo methods with rejection sampling could alternatively be used to approximate these expectations by simulation.

B Louis' Method for Estimates Obtained by Simultaneous Maximization

Here we describe how to implement Louis' (Louis, 1982) method based on the identity

$$I_{\text{OBS}}(\theta) = \sum_{i=1}^{m} E[I_{\text{C}i}(\theta)|D_i] - \sum_{i=1}^{m} E[S_i(\theta)S_i'(\theta)|D_i] + \sum_{i=1}^{m} E[S_i(\theta)|D_i]\{E[S_i(\theta)|D_i]\}' \ . \quad \text{(B.1)}$$

For simplicity, hereafter we drop the subscript i and only consider a single observation. Then the complete data score function, obtained from (3.3), is

$$S(\theta) = (S_1'(\theta_1), S_2'(\theta_2))' \ , \quad \text{(B.2)}$$

where $S_1(\theta_1) = (S'_{11}(\theta_1), S'_{12}(\theta_1))'$ with $S_{11}(\theta_1) = \partial \log L_{\rm C}(\theta)/\partial\alpha_1$ and $S_{12}(\theta_1) = \partial \log L_{\rm C}(\theta)/\partial\beta_1$, and $S_2(\theta_2) = (S'_{21}(\theta_2), S'_{22}(\theta_2))'$ with $S_{21}(\theta_2) = \partial \log L_{\rm C}(\theta)/\partial\alpha_2$ and $S_{22}(\theta_2) = \partial \log L_{\rm C}(\theta)/\partial\beta_2$. The corresponding contribution to the complete data information matrix is then

$$I_i = -\begin{pmatrix} \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_1\partial\alpha_1'} & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_1\partial\beta_1'} & 0 & 0 \\ \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_1\partial\alpha_1'} & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_1\partial\beta_1'} & 0 & 0 \\ 0 & 0 & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_2\partial\alpha_2'} & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_2\partial\beta_2'} \\ 0 & 0 & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_2\partial\alpha_2'} & \dfrac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_2\partial\beta_2'} \end{pmatrix}.$$

For the specific models we consider,

$$\frac{\partial \log L_{\rm C}(\theta)}{\partial\alpha_{1k}} = \delta_1\left\{\frac{I_k(t_1)}{\alpha_{1k}} - W_k(t_1)e^{x'\beta_1}\right\} - (1-\delta_1)W_k(a_{R_1})e^{x'\beta_1}\ , \quad k = 1, 2, \ldots, K_1\ ,$$

$$\frac{\partial \log L_{\rm C}(\theta)}{\partial\beta_1} = \delta_1\left\{\sum_{k=1}^{K_1}\left(I_k(t_1) - \alpha_{1k}W_k(t_1)e^{x'\beta_1}\right)\right\}x - (1-\delta_1)\left\{\sum_{k=1}^{K_1}\alpha_{1k}W_k(a_{R_1})e^{x'\beta_1}\right\}x\ ,$$

$$\frac{\partial \log L_{\rm C}(\theta)}{\partial\alpha_{2k}} = \delta_1\sum_{r=1}^{R_2}\left(\frac{n_{rk}}{\alpha_{2k}} - u_{rk}e^{x'\beta_2}\right) = \delta_1\sum_{r=1}^{R_2}\left(\frac{n_r u_{rk}}{\sum_{j=1}^{K_2}\alpha_{2j}u_{rj}} - u_{rk}e^{x'\beta_2}\right)\ , \quad k = 1, 2, \ldots, K_2\ ,$$

$$\frac{\partial \log L_{\rm C}(\theta)}{\partial\beta_2} = \delta_1\sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\left(n_{rk} - \alpha_{2k}u_{rk}e^{x'\beta_2}\right)x = \delta_1\left(n - \sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\alpha_{2k}u_{rk}e^{x'\beta_2}\right)x\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_{1j}\partial\alpha_{1k}} = 0\ , \quad j \neq k\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_{1k}^2} = \delta_1\frac{I_k(t_1)}{\alpha_{1k}^2}\ , \quad k = 1, 2, \ldots, K_1\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_1\partial\alpha_{1k}} = \delta_1 W_k(t_1)e^{x'\beta_1}x + (1-\delta_1)W_k(a_{R_1})e^{x'\beta_1}x\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_1\partial\beta_1'} = \delta_1\left\{\sum_{k=1}^{K_1}\alpha_{1k}W_k(t_1)e^{x'\beta_1}\right\}xx' + (1-\delta_1)\left\{\sum_{k=1}^{K_1}\alpha_{1k}W_k(a_{R_1})e^{x'\beta_1}\right\}xx'\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\alpha_{2k}\partial\alpha_{2\ell}} = \delta_1\sum_{r=1}^{R_2}\frac{n_r u_{rk}u_{r\ell}}{\left(\sum_{j=1}^{K_2}\alpha_{2j}u_{rj}\right)^2}\ , \quad k, \ell = 1, 2, \ldots, K_2\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_2\partial\alpha_{2k}} = \delta_1\sum_{r=1}^{R_2}\sum_{k=1}^{K_2}u_{rk}e^{x'\beta_2}x\ ,$$

$$-\frac{\partial^2 \log L_{\rm C}(\theta)}{\partial\beta_2\partial\beta_2'} = \delta_1\sum_{r=1}^{R_2}\sum_{k=1}^{K_2}\alpha_{2k}u_{rk}e^{x'\beta_2}xx'\ .$$

The conditional expectations can be evaluated by Gaussian Quadrature described above and based on the conditional probability density function (3.7). And to obtain the variance estimate, we need the inverse of the observed information matrix, thus in this case, we need to solve a high dimensional

matrix. Instead of using the *solve* function, we used *ginv* function in *MASS* library (or *chol2inv* function).

C VARIANCE ESTIMATION FOLLOWING TWO-STAGE ESTIMATION

We estimate the asymptotic covariance matrix in the spirit of parametric two-stage estimation procedure; see Newey and McFadden (Newey and McFadden, 1994). The complete data score functions are shown in (B.2). For the simultaneous estimation approach, we solve the following estimating functions:

$$U(\theta) = 0 \ ,$$

where

$$U(\theta) = \begin{pmatrix} U_{11}(\theta_1,\theta_2) \\ U_{12}(\theta_1,\theta_2) \\ U_{21}(\theta_1,\theta_2) \\ U_{22}(\theta_1,\theta_2) \end{pmatrix} = \begin{pmatrix} E\left[S_{11}(\theta_1)|D;\theta_1,\theta_2\right] \\ E\left[S_{12}(\theta_1)|D;\theta_1,\theta_2\right] \\ E\left[S_{21}(\theta_2)|D;\theta_1,\theta_2\right] \\ E\left[S_{22}(\theta_2)|D;\theta_1,\theta_2\right] \end{pmatrix} \ . \tag{C.1}$$

For the two-stage estimation approach, in the first stage we solve

$$U_1^*(\theta_1) = \begin{pmatrix} U_{11}^*(\theta_1) \\ U_{12}^*(\theta_1) \end{pmatrix} = \begin{pmatrix} E\left[S_{11}(\theta_1)|\mathcal{C}_1,X;\theta_1\right] \\ E\left[S_{12}(\theta_1)|\mathcal{C}_1,X;\theta_1\right] \end{pmatrix} = 0 \ , \tag{C.2}$$

and in the second stage we solve

$$U_2^*(\theta_2) = \begin{pmatrix} U_{21}^*(\theta_2) \\ U_{22}^*(\theta_2) \end{pmatrix} = \begin{pmatrix} E[S_{21}(\theta_2)|D;\widehat{\theta}_1,\theta_2] \\ E[S_{22}(\theta_2)|D;\widehat{\theta}_1,\theta_2] \end{pmatrix} = 0 \ . \tag{C.3}$$

Thus at the second stage of the two-stage procedure we plug $\widehat{\theta}_1$ into (C.1) and estimate θ_2 by solving the resulting equation. Let $U_1(\theta_1,\theta_2) = (U_{11}'(\theta_1,\theta_2), U_{12}'(\theta_1,\theta_2))'$, $U_2(\theta_1,\theta_2) = (U_{21}'(\theta_1,\theta_2), U_{22}'(\theta_1,\theta_2))'$, and $U_2^*(\theta_2) = U_2(\widehat{\theta}_1,\theta_2)$. Then if $\theta_0 = (\theta_{10}',\theta_{20}')'$ denotes the true value of θ, consider the Taylor expansion of the score function $U_2^*(\theta_2)$ around θ_0 and evaluate it at $\widehat{\theta}_2$ giving,

$$0 = U_2^*(\widehat{\theta}_2) = U_2(\widehat{\theta}_1,\widehat{\theta}_2) = U_2(\theta_{10},\theta_{20}) + \frac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_1}\left(\widehat{\theta}_1-\theta_{10}\right) + \frac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_2}\left(\widehat{\theta}_2-\theta_{20}\right) + o_p(n^{1/2}) \ .$$

Also

$$0 = U_1^*(\widehat{\theta}_1) = U_1^*(\theta_{10}) + \frac{\partial U_1^*(\theta_{10})}{\partial\theta_1}\left(\widehat{\theta}_1-\theta_{10}\right) + o_p(n^{1/2}) \ ,$$

therefore,

$$\begin{pmatrix} U_1^*(\theta_{10}) \\ U_2(\theta_{10},\theta_{20}) \end{pmatrix} = -\begin{pmatrix} \dfrac{\partial U_1^*(\theta_{10})}{\partial\theta_1} & 0 \\ \dfrac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_1} & \dfrac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_2} \end{pmatrix} \begin{pmatrix} \widehat{\theta}_1-\theta_{10} \\ \widehat{\theta}_2-\theta_{20} \end{pmatrix} \ .$$

As $n\to\infty$, by the law of large numbers

$$-\frac{1}{n}\begin{pmatrix} \dfrac{\partial U_1^*(\theta_{10})}{\partial\theta_1} & 0 \\ \dfrac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_1} & \dfrac{\partial U_2(\theta_{10},\theta_{20})}{\partial\theta_2} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} I_{11}^* & 0 \\ I_{21} & I_{22} \end{pmatrix} \triangleq A \ ,$$

and by the central limit theorem,

$$\frac{1}{\sqrt{n}} \begin{pmatrix} U_1^*(\theta_{10}) \\ U_2(\theta_{10}, \theta_{20}) \end{pmatrix} \xrightarrow{d} N(0, B) \ , \quad \text{where } B = \begin{pmatrix} I_{11}^* & 0 \\ 0 & I_{22} \end{pmatrix} .$$

Therefore,

$$\sqrt{n} \begin{pmatrix} \widehat{\theta}_1 - \theta_{10} \\ \widehat{\theta}_2 - \theta_{20} \end{pmatrix} \xrightarrow{d} N(0, A^{-1}BA^{-1'}) .$$

ACKNOWLEDGEMENTS

The authors thank Dr Dafna Gladman and Dr Vinod Chandran for stimulating collaboration and helpful discussions involving the psoriatic arthritis research program.

FUNDING

Natural Sciences and Engineering Research Council of Canada (RGPIN 155849); Canadian Institutes for Health Research (FRN 13887); Canada Research Chair (Tier 1) - CIHR funded (950-226626).

REFERENCES

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). *Statistical Models Based on Counting Processes*. Springer Science & Business Media, New York.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Barlow, R. E., Bartholomew, D. J., Bremner, J., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley New York.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., and Gladman, D. D. (2010). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7):1399–1405.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79(387):632–638.

Golub, G. H. and Welsch, J. H. (1969). Calculation of gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56(296):841–868.

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23(9):1455–1497.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Martinussen, T. and Scheike, T. H. (2007). *Dynamic Regression Models for Survival Data*. Springer Science & Business Media.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Rahman, P., Gladman, D. D., Cook, R. J., Zhou, Y., Young, G., and Salonen, D. (1998). Radiological assessment in psoriatic arthritis. *Rheumatology*, 37(7):760–765.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Wang, H. and Leng, C. (2008). A note on adaptive group LASSO. *Computational Statistics and Data Analysis*, 52(12):5277–5286.

Wu, Y. and Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics*, 71:782–791.

Yuan, M. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67.